

# Addressing moral problems through practical reasoning

Katie Atkinson \*, Trevor Bench-Capon

*Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK*

Available online 29 June 2007

---

## Abstract

In this paper, following the work of Hare, we consider moral reasoning not as the application of moral norms and principles, but as reasoning about what ought to be done in a particular situation, with moral norms perhaps emerging from this reasoning. We model this situated reasoning drawing on our previous work on argumentation schemes, here set in the context of Action-Based Alternating Transition Systems. We distinguish what prudentially ought to be done from what morally ought to be done, consider what legislation might be appropriate and characterise the differences between morally correct, morally praiseworthy and morally excusable actions. We also describe an implementation which simulates this reasoning and discuss some issues arising from the simulation.

© 2007 Elsevier B.V. All rights reserved.

**Keywords:** Practical reasoning; Argumentation; Values; Audiences; Moral principles; AATS

---

## 1. Introduction

In *Freedom and Reason* [16], R.M. Hare, perhaps the leading British moral philosopher of the twentieth century, notes that:

“There is a great difference between people in respect of their readiness to qualify their moral principles in new circumstances. One man may be very hidebound: he may feel that he knows what he ought to do in a certain situation as soon as he has acquainted himself with its most general features. . . . Another man may be more cautious. . . . he will never make up his mind what he ought to do, even in a quite familiar situation, until he has scrutinised every detail.” (p. 41)

Hare regards both these extreme positions as incorrect:

“What the wiser among us do is to think deeply about the crucial moral questions, especially those that face us in our own lives, but when we have arrived at an answer to a particular problem, to crystallise it into a not too specific or detailed form, so that its salient features may stand out and serve us again in a like situation without so much thought.” (pp. 41–42)

---

\* Corresponding author.

E-mail addresses: [katie@csc.liv.ac.uk](mailto:katie@csc.liv.ac.uk) (K. Atkinson), [tbc@csc.liv.ac.uk](mailto:tbc@csc.liv.ac.uk) (T. Bench-Capon).

Thus, for Hare, while everyday moral decisions may be made by applying principles and norms, serious moral decisions require reasoning about the particular situation, and it is such reasoning that gives rise to the moral principles that we subsequently apply. Moral norms are thus an output from, not an input to, serious moral reasoning. In this paper we will try to model such serious moral reasoning, with a view to making it properly computational so that autonomous software agents can engage in this form of reasoning. In doing so we will distinguish at least three things that might be intended by “agent A should  $\phi$ ”. We might mean something like “it is prudent to  $\phi$ ”, as when we say “you should wear a coat when the weather is cold”. Here the obligation is determined only by reference to the interests of the agent doing the reasoning. Alternatively, we might mean “it is morally right to  $\phi$ ”, as when we say “you should tell the truth”. Here the obligation is required to reflect the interests not only of the reasoning agent, but also of other agents affected by the action. Thirdly, we might mean “it is legally obligated to  $\phi$ ” as in “you should pay your taxes”, where the obligation derives from a legal system with jurisdiction over the agent. We will explore the differences between these three senses of “should”: in particular we will explain the difference between prudential “should” and moral “should” in terms of the practical reasoning involved, and consider the reasoning that might be used in devising appropriate legislation.

We will base our considerations on the representation and discussion of a specific example, a well known problem intended to explore a particular ethical dilemma discussed by Coleman [13] and Christie [12], amongst others. The situation involves two agents, called Hal and Carla, both of whom are diabetic. Hal, *through no fault of his own*, has lost his supply of insulin and urgently needs to take some to stay alive. Hal is aware that Carla has some insulin kept in her house, but Hal does not have permission to enter Carla’s house. The question is whether Hal is justified in breaking into Carla’s house and taking her insulin in order to save his life. Note that by taking Carla’s insulin, Hal may be putting her life in jeopardy, since she will come to need that insulin herself. One possible response is that if Hal has money, he can compensate Carla so that her insulin can be replaced before she needs it. Alternatively if Hal has no money but Carla does, she can replace her insulin herself, since her need is not immediately life threatening. There is, however, a serious problem if neither have money, since in that case Carla’s life is really under threat. Coleman argued that Hal may take the insulin to save his life, but should compensate Carla. Christie’s argument against this was that even if Hal had no money and was unable to compensate Carla he would still be justified in taking the insulin by his immediate necessity, since no one should die because of poverty. Thus, argues Christie, Hal cannot be *obliged* to compensate Carla even when he is able to.

In Section 2, we model our agents as simple automata and describe Action-Based Alternating Transition Systems (AATS) [22], which we use as the semantic basis of our representation, and instantiate an AATS relevant to the problem scenario. In any particular situation, the agents will need to choose how to act. In Section 3 we model this choice as the proposal, critique and defence of arguments justifying their available actions in the manner of [8]. In Section 4 we show how reasoning about the resulting arguments can be represented as a Value-Based Argumentation Framework [9] to enable the agents to identify strategies that are prudentially and morally justified. In Section 5 we consider how this framework can also be used to answer the question of what would be appropriate legislation for the situation, and what could be appropriate moral principles to take from the reasoning. Section 6 discusses some empirical simulations based on this approach and Section 7 concludes the paper.

## 2. Representing the problem

In this section we will describe a representation of the problem. Here we will follow the representation used in [4]. Building an implementation of the approach to allow simulations to be run [11] did lead to some refinements, required to block spurious arguments from being generated. For example, as given here, it appears that Hal can buy insulin instead of taking Carla’s, even though this is excluded in the problem statement. For the simulation in [11] an additional attribute was included in the state to indicate whether or not insulin was available for purchase. In order to keep things simple, however, we will not include these refinements in this section: they will be further discussed in Section 6.

In the simpler representation three attributes of agents are important: whether they have insulin (I), whether they have money (M) and whether they are alive (A). The state of an agent may thus be represented as a vector of three digits, IMA, with I, M and A equal to 1 if the agent has insulin, has money and is alive, respectively, and 0 if these things are false. Since I cannot be true and A false (the agent will live if it has insulin), an agent may be in any one of six possible states. We may now represent the actions available to the agents by depicting them as automata, as shown

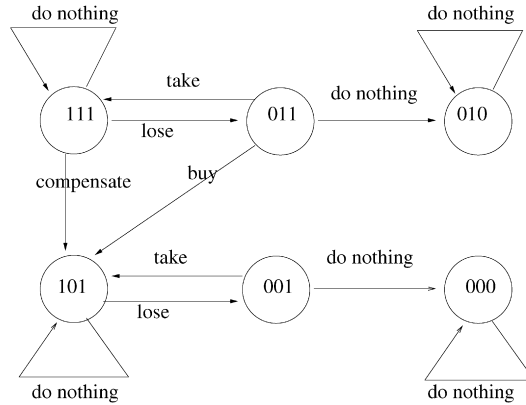


Fig. 1. State transition diagram for our agents.

in Fig. 1. An agent with insulin may lose its insulin; an agent with money and insulin may compensate another agent; an agent with no insulin may take another's insulin, or, with money, buy insulin. In any situation when it is alive, an agent may choose to do nothing; if dead it can only do nothing.

Next we draw upon the approach of Wooldridge and van der Hoek [22] which formally describes a normative system in terms of constraints on actions that may be performed by agents in any given state. We will now briefly summarise their approach.

In [22] Wooldridge and van der Hoek present an extension to Alur et al.'s Alternating-time Temporal Logic (ATL) [1], Normative ATL\* (NATL\*). As Wooldridge and van der Hoek explain, ATL is a logic of cooperative ability. Its purpose is to support reasoning about the powers of agents and coalitions of agents in game-like multi-agent systems. ATL contains an explicit notion of agency, which gives it the flavour of an action logic. NATL\* is intended to provide a link between ATL and deontic logic and the work presented in [22] provides a formal model to represent the relationship between agents' abilities and obligations. The semantic structures that underpin ATL are known as *Action-Based Alternating Transition Systems* (AATSs) and they are used for modelling game-like, dynamic, multi-agent systems. Such systems comprise multiple agents that can perform actions in order to modify and attempt to control the system in some way. In Wooldridge and van der Hoek's approach they use an AATS to model the physical properties of the system in question—the actions that agents can perform in the empty normative system, without regard to their legality or usefulness. They define an AATS as follows.

Firstly the systems of interest may be in any of a finite set  $Q$  of possible *states*, with some  $q_i \in Q$  designated as the *initial state*. Systems are populated by a set  $Ag$  of *agents*; a *coalition* of agents is simply a set  $C \subseteq Ag$ , and the set of all agents is known as the *grand coalition*. Note, Wooldridge and van der Hoek's usage of the term 'coalition' does not imply any common purpose or shared goal: a coalition is simply taken to be a set of agents.

Each agent  $i \in Ag$  is associated with a set  $Ac_i$  of possible actions, and it is assumed that these sets of actions are pairwise disjoint (i.e., actions are unique to agents). The set of actions associated with a coalition  $C \subseteq Ag$  is denoted by  $Ac_C$ , so  $Ac_C = \bigcup_{i \in C} Ac_i$ .

A joint action  $j_C$  for a coalition  $C$  is a tuple  $\langle \alpha_1, \dots, \alpha_k \rangle$ , where for each  $\alpha_j$  (where  $j \leq k$ ) there is some  $i \in C$  such that  $\alpha_j \in Ac_i$ . Moreover, there are no two different actions  $\alpha_j$  and  $\alpha_{j'}$  in  $j_C$  that belong to the same  $Ac_i$ . The set of all joint actions for coalition  $C$  is denoted by  $J_C$ , so  $J_C = \prod_{i \in C} Ac_i$ . Given an element  $j$  of  $J_C$  and an agent  $i \in C$ ,  $i$ 's component of  $j$  is denoted by  $j_i$ .

An *Action-Based Alternating Transition System* (AATS) is an  $(n + 7)$ -tuple  $S = \langle Q, q_0, Ag, Ac_1, \dots, Ac_n, \rho, \tau, \Phi, \pi \rangle$ , where:

- $Q$  is a finite, non-empty set of *states*;
- Some  $q_i \in Q$  is designated as the *initial state*;
- $Ag = \{1, \dots, n\}$  is a finite, non-empty set of *agents*;
- $Ac_i$  is a finite, non-empty set of *actions*, for each  $i \in Ag$  where  $Ac_i \cap Ac_j = \emptyset$  for all  $i \neq j \in Ag$ ;

- $\rho : Ac_{Ag} \rightarrow 2^Q$  is an *action precondition function*, which for each action  $\alpha \in Ac_{Ag}$  defines the set of states  $\rho(\alpha)$  from which  $\alpha$  may be executed;
- $\tau : Q \times J_{Ag} \rightarrow Q$  is a partial *system transition function*, which defines the state  $\tau(q, j)$  that would result by the performance of  $j$  from state  $q$ —note that, as this function is partial, not all joint actions are possible in all states (cf. the precondition function above);
- $\Phi$  is a finite, non-empty set of *atomic propositions*; and
- $\pi : Q \rightarrow 2^\Phi$  is an interpretation function, which gives the set of primitive propositions satisfied in each state: if  $p \in \pi(q)$ , then this means that the propositional variable  $p$  is satisfied (equivalently, true) in state  $q$ .

In addition to the elements of an AATS given here, we need to provide an extension to enable the representation of values<sup>1</sup> from the underlying theory of practical reasoning, as will be presented in the next section. Firstly, we have a set  $Av$  of values for each agent (which are a subset of a set  $V$  of values). Every transition between two states from the set  $Q$  either promotes, demotes, or is neutral, with respect to each value. Note that values are not unique to agents: individual agents may or may not have values in common. Whether a value is promoted or demoted by a given action will be determined by comparing the state reached with the state left. More formal definitions of these elements are given below:

- $Av_i$  is a finite, non-empty set of values  $Av_i \subseteq V$ , for each  $i \in Ag$ .
- $\delta : Q \times Q \times Av_{Ag} \rightarrow \{+, -, =\}$  is a *valuation function* which defines the status (promoted (+), demoted (−) or neutral (=)) of a value  $v_u \in Av_{Ag}$  ascribed by the agent to the transition between two states:  $\delta(q_x, q_y, v_u)$  labels the transition between  $q_x$  and  $q_y$  with one of  $\{+, -, =\}$  with respect to the value  $v_u \in Av_{Ag}$ .

We can now extend the original specification of an AATS to accommodate the notion of values and we re-define an AATS as a  $(2n + 8)$  tuple  $S = \langle Q, q_0, Ag, Ac_1, \dots, Ac_n, Av_1, \dots, Av_n, \rho, \tau, \Phi, \pi, \delta \rangle$ .

We now turn to representing the insulin scenario as an AATS. Recall from Section 2 that each agent may independently be in one of six states, giving 36 possible states for the two agents,  $q_0 \dots q_{35}$ . Normally both agents will have insulin, but we are specifically interested in the situations that arise when one of them (Hal) loses his insulin. The initial state therefore may be any of the four states in which  $I_H = 0$ . Moreover, since Hal is supposed to have no time to buy insulin, his only available actions in these states, whether or not  $M_H = 1$ , are to take Carla's insulin or do nothing. If Hal does nothing, neither agent can act further. If Hal takes Carla's insulin and if  $M_H = 1$ , then Hal can compensate Carla or do nothing. Similarly, after Hal takes the insulin, Carla, if  $M_C = 1$ , can buy insulin or do nothing. The possible developments from the four initial states are shown in Fig. 2. States are labelled with the two vectors  $I_H M_H A_H$  (on the top row) representing Hal's situation and  $I_C M_C A_C$  (on the bottom row) representing Carla's situation. The arcs are labelled with the joint actions (with the other labels on the arcs to be explained in Section 4).

The instantiation of the problem as an AATS is summarised below. We give only the joint actions and the transitions for scenarios where Hal has lost his insulin.

*States and initial states:*

$$Q = \{q_0, \dots, q_{35}\}.$$

As shown in Fig. 2, the initial state is one of four,  $\{q_0, q_1, q_2, q_3\}$ .

*Agents, actions and joint actions:*

$$\begin{aligned} Ag &= \{H, C\}, & Ac_H &= \{\text{take}_H, \text{compensate}_H, \text{do\_nothing}_H\}, & Ac_C &= \{\text{buy}_C, \text{do\_nothing}_C\}, \\ J_{AG} &= \{j_0, j_1, j_2, j_3\}, & \text{where } j_0 &= \langle \text{do\_nothing}_H, \text{do\_nothing}_C \rangle, & j_1 &= \langle \text{take}_H, \text{do\_nothing}_C \rangle, \\ & & j_2 &= \langle \text{do\_nothing}_H, \text{buy}_C \rangle, & j_3 &= \langle \text{compensate}_H, \text{do\_nothing}_C \rangle. \end{aligned} \quad ^2$$

<sup>1</sup> Values denote some actual descriptive social attitude/interest which the proponent of an action may or may not wish to uphold or subscribe to. Note that these are descriptions such as 'life', 'liberty' and 'happiness', and not numeric quantities. Values and their distinction from goals are discussed in more detail in Section 3.

<sup>2</sup> Here we take 'compensate' to mean that Hal buys insulin and gives it to Carla or, if Carla already has insulin, Hal gives her money.

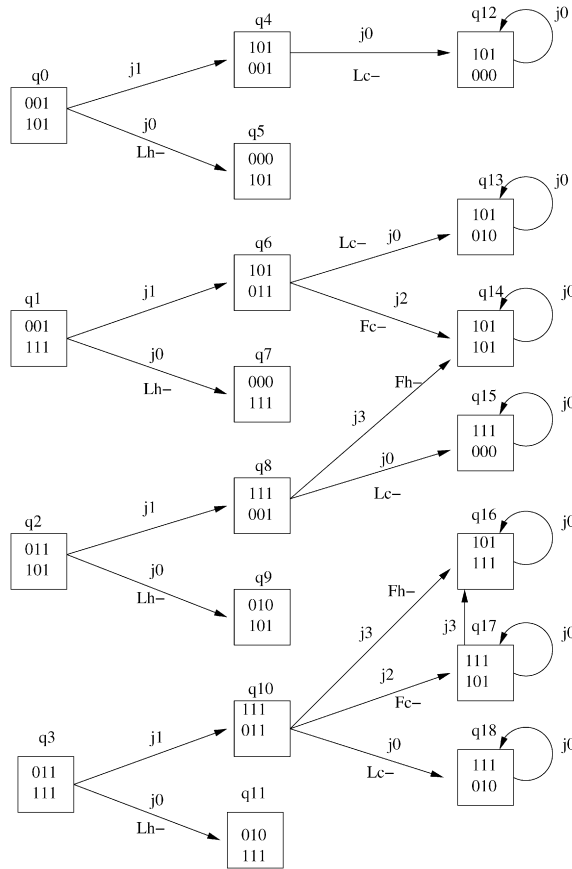


Fig. 2. Developments from the four possible initial states.

Table 1  
Preconditions/Transitions/Interpretations of preconditions

$q \backslash j$	j0	j1	j2	j3	$\pi(q)$
q0	q5	q4	–	–	{ $alive_H, insulin_C, alive_C$ }
q1	q7	q6	–	–	{ $alive_H, insulin_C, money_C, alive_C$ }
q2	q9	q8	–	–	{ $money_H, alive_H, insulin_C, alive_C$ }
q3	q11	q10	–	–	{ $money_H, alive_H, insulin_C, money_C, alive_C$ }
q4	q12	–	–	–	{ $insulin_H, alive_H, alive_C$ }
q5	–	–	–	–	{ $insulin_C, alive_C$ }
q6	q13	–	q14	–	{ $insulin_H, alive_H, money_C, alive_C$ }
q7	–	–	–	–	{ $insulin_C, money_C, alive_C$ }
q8	q15	–	–	q14	{ $insulin_H, money_H, alive_H, alive_C$ }
q9	–	–	–	–	{ $money_H, insulin_C, alive_C$ }
q10	q18	–	q17	q16	{ $insulin_H, money_H, alive_H, money_C, alive_C$ }
q11	–	–	–	–	{ $money_H, insulin_C, money_C, alive_C$ }
q12	q12	–	–	–	{ $insulin_H, alive_H$ }
q13	q13	–	–	–	{ $insulin_H, alive_H, money_C$ }
q14	q14	–	–	–	{ $insulin_H, alive_H, insulin_C, alive_C$ }
q15	q15	–	–	–	{ $insulin_H, money_H, alive_H$ }
q16	q16	–	–	–	{ $insulin_H, alive_H, insulin_C, money_C, alive_C$ }
q17	q17	–	–	q16	{ $insulin_H, money_H, alive_H, insulin_C, alive_C$ }
q18	q18	–	–	–	{ $insulin_H, money_H, alive_H, money_C$ }

*Propositional variables:*

$$\Phi = \{\text{insulin}_H, \text{money}_H, \text{alive}_H, \text{insulin}_C, \text{money}_C, \text{alive}_C\}.$$

*Preconditions/Transitions/Interpretations of preconditions for states used in the scenario are given in Table 1.*

### 3. Constructing the arguments

In [7] we proposed an argument scheme and associated critical questions to enable agents to propose, attack and defend justifications for action. Such an argument scheme follows Walton [21] in viewing reasoning about action (practical reasoning) as presumptive justification—*prima facie* justifications of actions can be presented as instantiations of an appropriate argument scheme, and then critical questions characteristic of the scheme used can be posed to challenge these justifications. The argument scheme we have developed is an extension of Walton’s *sufficient condition scheme for practical reasoning* [21] and our argument scheme is stated as follows:

AS1 In the current circumstances R  
 We should perform action A  
 Which will result in new circumstances S  
 Which will realise goal G  
 Which will promote some value V.

This extended scheme can be viewed as a ‘positive’ justification as it proposes an action that would promote some value which is desirable for the agent. Additionally, AS1 can also be stated in a ‘negative’ version. Where a value is demoted we have a reason to refrain from an action to avoid a state containing particular features (a “goal”) which demote a value. This negative version, AS2, is stated as follows:

AS2 In the current circumstances R  
 We should perform action A  
 To avoid some circumstances S  
 Which would realise goal G  
 Which would demote some value V.

This negative version of AS1 can thus be used to argue in terms of avoiding some undesirable outcome, rather than achieving some positive outcome. The importance of arguments for avoiding undesirable outcomes has received emphasis from Amgoud and her colleagues (e.g. [2]), who use the term ‘bipolar goals’ to distinguish those to be pursued from those to be avoided. We relate this bipolarity not to goals (in some cases a goal which demotes one value may promote another, and so goals need not be intrinsically negative or positive), but to demotion or promotion of particular values. We make use of this negative variant of the argument scheme to generate arguments and critical questions for our example given later in this section, but once the arguments have been generated we make no distinction between those based on demotion rather than promotion of values.

In both our extended schemes we have refined Walton’s notion of a goal by distinguishing three elements: the state of affairs brought about by the action; the goal (the desired features in that state of affairs); and the value (the reason why those features are desirable). Our underlying idea in making this distinction is that the agent performs an action to move from one state of affairs to another. The new state of affairs may have many differences from the current state of affairs, and it may well be that only some of them are significant to the agent. The significance of these differences is that they make the new state of affairs better with respect to some good valued by the agent. Any of our three elements, S, G and V, can be used as the goal in Walton’s formulation. The appropriate questions and their resolution, however, can differ in the three cases. Note, with regard to the goods valued by agents, that typically the new state of affairs will be better through improving the lot of some *particular* agent: the sum of human happiness is increased only by increasing the happiness of some particular human. In this paper we take the common good of all agents as the aggregation of their individual goods. It may be that there are common goods which are not reflected in this aggregation: for example, if equality is such a common good, increasing the happiness of an already happy agent may diminish the overall common good. For simplicity, we ignore such possibilities here.

Now an agent who does not accept a presumptive argument may attack the contentious elements in the instantiation through the application of critical questions. We have elaborated Walton's original four critical questions associated with his scheme by extending them to address the different elements identified in the goal in our new argument scheme. Our original extension resulted in sixteen different critical questions, which are as follows:

- CQ1: Are the believed circumstances true?
- CQ2: Assuming the circumstances, does the action have the stated consequences?
- CQ3: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal?
- CQ4: Does the goal realise the value stated?
- CQ5: Are there alternative ways of realising the same consequences?
- CQ6: Are there alternative ways of realising the same goal?
- CQ7: Are there alternative ways of promoting the same value?
- CQ8: Does doing the action have a side effect which demotes the value?
- CQ9: Does doing the action have a side effect which demotes some other value?
- CQ10: Does doing the action promote some other value?
- CQ11: Does doing the action preclude some other action which would promote some other value?
- CQ12: Are the circumstances as described possible?
- CQ13: Is the action possible?
- CQ14: Are the consequences as described possible?
- CQ15: Can the desired goal be realised?
- CQ16: Is the value indeed a legitimate value?

Thus, in posing such critical questions agents can attack the validity of the various elements of the argument scheme and the connections between them, and additionally there may be alternative possible actions, and side effects of the proposed action. Each critical question can be seen as an attack on the argument it is posed against, as described in [3]. However, since the representation method that we use in this paper involves agents participating in joint actions, there is one further critical question that arises here:

- CQ17: Is the other agent guaranteed to execute its part of the desired joint action?

In [3] this objection was subsumed under CQ2 since it also claims that the action will lead to a different state from that proposed in the argument. There is, however, a point in distinguishing cases where the different state is reached through the choice of another agent, since the method of rebuttal will differ. For example, we may answer this critical question by explaining why it is in the interests of the other agent to participate in this joint action. We therefore add this additional critical question as CQ17. It becomes of particular importance in the empirical investigations discussed in Section 6.

To summarise, we therefore believe that in an argument about a matter of practical action, we should expect to see one or more *prima facie* justifications advanced stating, explicitly or implicitly, the current situation, an action, the situation envisaged to result from the action, the features of that situation for which the action was performed and the value promoted by the action, together with negative answers to critical questions directed at those claims. We now describe how this approach to practical reasoning can be represented in terms of an AATS. In this paper we present the arguments and critical questions in a semi-formal manner. A forthcoming paper [5] gives precise definitions of instantiations of AS1 and AS2 and all the associated critical questions. These definitions will enable general properties to be demonstrated, for example, through the use of model checking techniques similar to the approach to social laws given in [20].

In this particular scenario we recognise two values relative to each agent: life and freedom (the ability to act in a given situation). The value 'life' (L) is demoted when Hal or Carla ceases to be alive. The value 'freedom' (F) is demoted when Hal or Carla ceases to have money. The arcs in Fig. 2 are labelled with the value demoted by a transition, subscripted to show the agent in respect of which it is demoted. We can now examine the individual arguments involved.

In all of  $q_0 - q_3$ , the joint action  $j_0$  demotes the value ‘life’ in respect of Hal, whereas the action  $j_1$  is neutral with regard to this value. We can instantiate argument scheme AS2 by saying that where Hal has no insulin he should take Carla’s to avoid those states where dying demotes the value ‘life’.

A1: Where  $\text{Insulin}_h = 0$ ,  $\text{Take}_h$  (i.e.  $j_1$ ), To avoid  $\text{Alive}_h = 0$ , Which demotes  $L_h$ .

Argument A2 attacks A1 and it arises from  $q_0$  where Hal taking the insulin leads to Carla’s death and thus demotes the value ‘life Carla’. By ‘not take’ we mean any of the other available actions.

A2 attacks A1: Where  $\text{Money}_c = 0$ ,  $\text{Not Take}_h$  (i.e.  $j_0$  or  $j_2$ ), To avoid  $\text{Alive}_c = 0$ , Which demotes  $L_c$ .

Argument A3 arises from  $q_2$  where Carla’s death is avoided by Hal taking the insulin and paying Carla compensation.

A3 attacks A2 and A5: Where  $\text{Insulin}_h = 0$ ,  $\text{Take}_h$  and  $\text{Compensate}_h$  (i.e.  $j_1$  followed by  $j_3$ ), To achieve  $\text{Alive}_c = 1$  and  $\text{Money}_c = 1$ , Which promotes  $L_c$  and  $F_c$ .

Argument A4 represents a critical question directed at A2 which challenges the factual premise of A2, that Carla has no money.

A4 attacks A2:  $\text{Money}_c = 1$  (Known to Carla but not Hal).

Next argument A5 mutually attacks A3 and it also attacks A2. A5 states that where Hal has no insulin but he does have money, then he should take Carla’s insulin and she should buy some more. The consequences of this are that Carla remains alive, promoting the value ‘life Carla’, and, Hal has money, promoting the value ‘freedom Hal’.

A5 attacks A3 and A2: Where  $\text{Insulin}_h = 0$  and  $\text{Money}_h = 1$ ,  $\text{Take}_h$  and  $\text{Buy}_c$  (i.e.  $j_1$  followed by  $j_2$ ), To achieve  $\text{Alive}_c = 1$  and  $\text{Money}_h = 1$ , Which promotes  $L_c$  and  $F_h$ .

Argument A6 critically questions A5 by attacking the assumption in A5 that Carla has money.

A6 attacks A5:  $\text{Money}_c = 0$  (Known to Carla but not Hal).

Another attack on A5 can be made by argument A7 stating that where Carla has money then she should not buy any insulin so as to avoid having no money, which would demote the value ‘freedom Carla’.

A7 attacks A5: Where  $\text{Money}_c = 1$ ,  $\text{Not Buy}_c$  (i.e.  $j_0$  or  $j_1$  or  $j_3$ ), To avoid  $\text{Money}_c = 0$ , Which demotes  $F_c$ .

A8 is a critical question against A3 which states that where Hal does not have money, taking the insulin and compensating Carla is not a possible strategy.

A8 attacks A3: Where  $\text{Money}_h = 0$ ,  $\text{Take}_h$  and  $\text{Compensate}_h$  (i.e.  $j_1$  followed by  $j_3$ ), is not a possible strategy.

A8 is attacked by argument A9 which challenges the assumption in A8 that Hal has no money, and A9 is in turn attacked by A10 which challenges the opposite assumption, that Hal does have money.

A9 attacks A8 and A11:  $\text{Money}_h = 1$  (Known to Hal but not Carla).

A10 attacks A9:  $\text{Money}_h = 0$  (Known to Hal but not Carla).

Argument A11 attacks A1 in stating that where Hal does not have money but Carla does, then Hal should not take the insulin to avoid Carla being left with no money, which would demote the value ‘freedom Carla’.



A11 attacks A1: Where  $\text{Money}_h = 0$  and  $\text{Money}_c = 1$ , Not  $\text{Take}_h$  (i.e.  $j_0$ ), To avoid  $\text{Money}_c = 0$ , Which demotes  $F_c$ .

Argument A12 can attack A5 by stating that in the situations where Hal does not have insulin, then he should take Carla's insulin but not compensate her. This would avoid him being left with no money, as when Hal has no money the value 'freedom Hal' is demoted.

A12 attacks A5: Where  $\text{Insulin}_h = 0$ ,  $\text{Take}_h$  and Not  $\text{Compensate}_h$  (i.e.  $j_1$  followed by  $j_0$  or  $j_2$ ), To avoid  $\text{Money}_h = 0$ , Which demotes  $F_h$ .

Finally, argument A13 attacks A2 by stating that where Hal has no insulin and no money he should take Carla's insulin and she should buy some. This would ensure that Carla stays alive, promoting the value 'life Carla'.

A13 attacks A2: Where  $\text{Insulin}_h = 0$  and  $\text{Money}_h = 0$ ,  $\text{Take}_h$  and  $\text{Buy}_c$  (i.e.  $j_1$  followed by  $j_2$ ), To achieve  $\text{Alive}_c = 1$ , Which promotes  $L_c$ .

This concludes the description of the arguments and attacks that can be made by instantiating argument schemes AS1 and AS2, and posing appropriate critical questions.

#### 4. Evaluating the arguments

In the previous section we identified the arguments that the agents in our problem situation need to consider. In order to evaluate the arguments and see which ones the agents will accept, we organise the arguments into a Value-Based Argumentation Framework (VAF) [9]. VAFs extend the Argumentation Frameworks introduced by Dung in [14], so as to accommodate different audiences with different values and interests. One of the key notions in Dung's argumentation framework is that of a preferred extension (PE), a subset of the arguments in the framework which:

- is conflict free, in that no argument in the PE attacks any other argument in the PE;
- is able to defend every argument in the PE against attacks from outside the extension, in that every argument outside the PE which attacks an argument in the PE is attacked by some argument in the PE;
- is maximal, in that no other argument can be added to the PE without either introducing a conflict or an argument that cannot be defended against outside attacks.

In a VAF strengths of arguments for a particular *audience* are compared with reference to the *values* to which they relate. An audience has a *preference order* on the values of the arguments, and an argument is only *defeated for that audience* if its value is not preferred to that of its attacker. We then replace the notion of attack in Dung's PE by the notion of *defeat for an audience* to get the *PE for that audience*. We represent the VAF as a directed graph, the vertices representing arguments and labelled with an argument identifier and the value promoted by the argument, and the edges representing attacks between arguments. Attacks arise from the process of critical questioning, as described in the previous section, not from an analysis of the arguments themselves. The values promoted by the arguments are identified in the instantiations of the argument scheme presented in the previous section. The VAF for our problem scenario is shown in Fig. 3. Note that two pairs of arguments, {A4, A6} and {A9, A10} relate to facts known only to Carla and Hal respectively. In order to bring these into a value-based framework, we ascribe the value "truth" to statements of fact, and as in [9], truth is given the highest value preference for all audiences, since while we can choose what we consider desirable, we are constrained by the facts to accept what is true.<sup>3</sup>

The questions posed by the problem scenario are whether Hal should take the insulin and whether Hal should compensate Carla. We answer these questions by finding the preferred extensions (PEs) of the framework for various audiences. Note that the PE may contain both arguments providing reasons for performing an action and for not performing it. The actions which will be chosen are those supported by *effective* arguments, that is, those which do

<sup>3</sup> Ascribing the value 'truth' is used here as a convenient device. A framework to handle the mixture of epistemic and practical reasoning in legal decisions is described in [6].

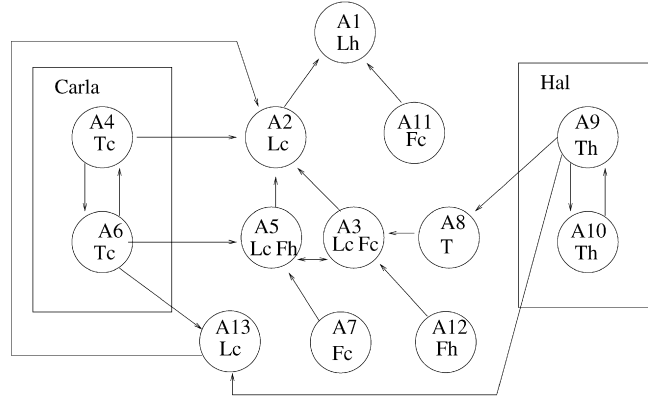


Fig. 3. VAF for the problem scenario.

not feature in an unsuccessful attack. Thus in  $q_0$ , for example, A2, which provides a reason for Hal not to take the insulin, is not attacked and so will be in the PE. If, however, we prefer  $L_H$  to  $L_C$ , A1, which gives a reason for Hal to take the insulin will also be included. In such a case A2 is ineffective and so Hal should take the insulin, despite there being reasons against this action which cannot be countered through argument. If A1 is in the PE it is always effective since it attacks nothing, and so if A1 is present then Hal should take the insulin. If A3, which gives a reason for Hal to compensate Carla, is included it is also always effective since it always defeats A2, because its values are a superset of A2, and it must defeat A5 or be defeated by it. If both A1 and A3 are in the PE, Hal should take the insulin and compensate Carla. If A3, but not A1, is present Hal should take the insulin *only* if he then compensates Carla. What we must do therefore is consider for which audiences A1 and A3 appear in the PE.

For this discussion we will assume that the agents are part of a common culture in which the value life is preferred to the value freedom. This seems reasonable in that life is a precondition for any exercise of freedom. There will therefore be no audience with a value order in which  $F_{Ag} > L_{Ag}$ , for any agent  $Ag$ , although of course it is possible for an agent to prefer its own freedom to the life of another.

First we note that  $\{A7, A11, A12\}$  are not attacked and so will appear in every PE. Immediately from this we see that A1 will not appear in any PE of an audience for which  $F_C \geq L_H$ , and that A3 will not appear in any PE of an audience for which  $F_H \geq L_C$ . A5 will never be defeated by A7, since  $L_C > F_C$  for all audiences.

To proceed we must now resolve the factual issues which determine the conflicts  $\{A4, A6\}$  and  $\{A9, A10\}$ . Thus we need to consider the initial states  $q_0 - q_3$  separately.

In states  $q_0$  and  $q_1$  A10 defeats A9 and hence A8 is included. Since truth is the highest value this will exclude A3 (reasonably enough since Hal *cannot* pay compensation). In  $q_0$  A6 defeats A4, A5 and A13, so that A2 is no longer attacked, and will be in the PE. In the presence of A2, we can include A1 only if  $L_H > L_C$ . Thus for  $q_0$  the PE will be  $\{A2, A6, A7, A8, A10, A11, A12\}$  extended with A1 for audiences for which  $L_H > L_C > F_C$ . In  $q_1$  A4 defeats A6 so A13 will be included. A4 also defeats A2 so A1 will be included for audiences for which  $L_H > F_C$ . Thus for  $q_1$  the PE will be  $\{A4, A13, A7, A8, A10, A11, A12\}$  extended with A1 for audiences for which  $L_H > F_C$ . In  $q_2$  and  $q_3$  A9 will defeat A10, A8 and A13. In  $q_2$  A6 defeats A4 and A5, so A3 will now be included for audiences for which  $L_C > F_H$ . If A3 is included A2 is defeated and A1 included, provided  $L_H > F_C$ . So the PE for  $q_2$  will be  $\{A6, A7, A9, A11, A12\}$  extended by A3 for audiences for which  $L_C > F_H$  and by A1 for audiences for which  $L_H > F_C$ . Finally in  $q_3$ , A4 defeats A6 and A2, so A1 is included if  $L_H > F_C$ . A5 and A3 are now in mutual conflict, and the conflict will be resolved depending on whether  $F_C$  or  $F_H$  is preferred. Thus the PE in  $q_3$  will contain  $\{A4, A7, A9, A11, A12\}$ , extended by A1 if  $L_H > F_C$ , by A3 if  $F_C > F_H$  and by A5 if  $F_H > F_C$ .

We can now summarise the status of A1 and A3 in Table 2.

From this we can see that if the interests of Hal are ranked above those of Carla, Hal should take the insulin and not pay compensation, whereas if the interests of Carla are ranked above those of Hal, Hal should take the insulin only if he compensates. These two positions thus express what is prudentially right for Hal and Carla to do respectively.

From the standpoint of pure morality, however, people should be treated equally: that is  $(L_H = L_C) > (F_H = F_C)$ . Remember, that if the problem is considered in the abstract, one does not know who will be the person who loses the insulin: one may find oneself playing the role of Hal or Carla, and so there is no reason to prefer one agent to the

Table 2  
Status of A1 and A3

Initial state	A1 included if:	A3 included if:
q <sub>0</sub>	$L_H > L_C > F_C$	never
q <sub>1</sub>	$L_H > F_C$	never
q <sub>2</sub>	$L_H > F_C$	$L_C > F_H$
q <sub>3</sub>	$L_H > F_C$	$F_C > F_H$
		A5 included otherwise

other. If this perspective is adopted, then Hal should take the insulin in all situations other than q<sub>0</sub>, and is obliged to compensate only in q<sub>2</sub>, since there are two PEs in q<sub>3</sub>. We can see this as representing the morally correct judgement, the judgement that would be arrived at by a neutral observer *in full possession of the facts*.

However, the point about being in full possession of the facts is important. In practice we need to evaluate the conduct of the agents in the situations in which they find themselves. In our scenario Hal cannot know whether or not Carla is in a position to replace the insulin herself: for Hal, q<sub>0</sub> is epistemically indistinguishable from q<sub>1</sub>, and q<sub>2</sub> is epistemically indistinguishable from q<sub>3</sub>. Now consider Hal in q<sub>2</sub>/q<sub>3</sub>. He will of course take the insulin and justify this by saying that his life is more important than Carla's freedom of choice with regard to her money. In a society which rates  $L > F$ , this will be accepted. Thus Hal should take the insulin. If he then chooses to compensate, he can be sure of acting in a morally acceptable manner, since this is required in q<sub>2</sub> and appears in one of the alternative PEs in q<sub>3</sub>. If, on the other hand, he does not compensate, while he may attempt justification in q<sub>3</sub> by saying that he saw no reason to prefer Carla's freedom of choice to his own, in q<sub>2</sub> he would have to argue that his freedom of choice is preferred to Carla's life. This justification will be rejected for the same reason that the justification for taking the insulin at all was accepted, namely that  $L > F$ . Morally, therefore, in q<sub>2</sub>/q<sub>3</sub>, Hal should take the insulin and compensate Carla.

Now consider q<sub>0</sub>/q<sub>1</sub>, where compensation is impossible. In q<sub>1</sub> taking the insulin is justifiable by  $L > F$ . In q<sub>0</sub>, however, the justification is only  $L_H > L_C$ . Hal's problem, if this is not acceptable, is that he cannot be sure of acting in a morally correct manner, since he could take the insulin in q<sub>1</sub> and not take it in q<sub>0</sub>. Our view is that taking the insulin should be seen as morally *excusable*, even in q<sub>0</sub>, although not morally *correct*, since the possibility of the actual state being q<sub>1</sub> at least excuses the preference of Hal's own interests to Carla's. The alternative is to insist on Hal not taking the insulin in q<sub>1</sub>, which could be explained only by  $L_H \leq F_C$ , and it seems impossibly demanding to expect Hal to prefer Carla's lesser interest to his own greater interest. The distinction between morally correct and morally excusable is merely our attempt to capture some of the nuances that are found in everyday discussions of right and wrong. There is considerable scope to explore these nuances, which are often obscured in standard deontic logic. See, for example, McNamara's discussion in [17] which distinguishes: what is *required* (what morality demands); what is *optimal* (what morality recommends); the supererogatory (exceeding morality's demands); the morally indifferent; the permissible suboptimal; the morally significant; and the minimum that morality demands. We will return to McNamara's concept of supererogation in Section 5.

There is also other work in the philosophy literature that makes use of the concept of values. For example, in [15] values are discussed in relation to a framework for reasoning about norms and preferences. However, this account differs from the one presented in this paper in its use of value statements. In [15] values are used in value judgements, e.g. by describing objects or states of affairs as being 'good' or 'bad'. Thus such value concepts can be used for classificatory purposes, e.g. in statements such as "This is a good car", and for comparative purposes, in statements such as "My new car is better than my old car". Although this account does relate states of affairs to values, it does so through the use of such value judgements, rather than by identifying context dependant values that are promoted through the attainment of goal states, as in the account presented in this paper.

## 5. Moral, prudential and legal "ought"

In our discussion in the previous section we saw that what an agent should do can be determined by the ordering the agent places on values. This ordering can take into account, or ignore, which of the agents the values relate to. Prudential reasoning takes account of the different agents, with the reasoning agent preferring values relating to itself,

whereas strict moral reasoning should ignore the individual agents and treat the values equally. We will consider five possible value orders which respect  $L > F$ , and which order the agents consistently.<sup>4</sup>

- V01 *Morally correct*: values are ordered – within each value agents are treated equally, and no distinctions relating to agents are made. In our example, for Hal:  $(L_H = L_C) > (F_H = F_C)$ .
- V02 *Self-interested*: values are ordered as for moral correctness, but within a value an agent prefers its own interests. In our example, for Hal:  $L_H > L_C > F_H > F_C$ .
- V03 *Selfish*: values are ordered, but an agent prefers its own interests to those of other agents: In our example, for Hal:  $L_H > F_H > L_C > F_C$ .
- V04 *Noble*: values are ordered as for moral correctness, but within a value an agent prefers the other's interests. In our example, for Hal:  $L_C > L_H > F_C > F_H$ .
- V05 *Sacrificial*: values are ordered, but an agent prefers the other's interests to its own. In our example, for Hal:  $L_C > F_C > L_H > F_H$ .

Note that the morally correct order is common to both agents, while the orders for self-interested Hal and noble Carla are the same, as are those for selfish Hal and sacrificial Carla.

Now in general an agent can determine what it should do by constructing the VAF comprising the arguments applicable in the situation and calculating the PE for that VAF using some value order. Using VO1 will give what it morally should do and VO3 what it prudentially should do.

It is, however, possible that there will not be a unique PE: this may be either because the value order cannot decide a conflict (as with A3 and A5 when using VO1 in  $q_3$  above), or because the agent lacks the factual information to resolve a conflict (as with Hal with respect to A4 and A6 above). In this case we need to consider all candidate PEs. In order to justify commitment to an action the agent will need to use a value order which includes the argument justifying the action in all candidate PEs.

Consider  $q_3$  and VO1: we have two PEs,  $\{A1, A3, A4, A7, A9, A11, A12\}$  and  $\{A1, A4, A5, A7, A9, A11, A12\}$ . A1 is in both and it is thus morally obligatory to take the insulin. A3 on the other hand is in one PE but not the other and so both compensate and not compensate are morally correct in  $q_3$ . It is possible to justify A5 by choosing a value order with  $F_H > F_C$ , or A3 by choosing a value order with  $F_C > F_H$ . Thus in  $q_3$  a selfish or a self-interested agent will not compensate, whereas a noble or sacrificing one will. Either choice is, however, consistent with the morally correct behaviour. Next we must consider what is known by the reasoning agent. Consider Hal in  $q_2/q_3$ , where we now have an additional PE to take into account. The PE from  $q_2$  using VO1 is  $\{A1, A3, A6, A7, A9, A11, A12\}$  and as A1 is in all three, taking the insulin is obligatory. To exclude A3 from the PE for  $q_2$ , the preference  $F_H > L_C$  is required so that the attack from A2 succeeds. Here legitimate self-interest cannot ground a choice not to compensate: this preference is only in VO3, which means that only a selfish agent will not compensate. In  $q_2$ , therefore, failing to compensate is not consistent with morally correct behaviour, and an agent which made this choice would be subject to moral condemnation. VO2 cannot exclude A3 from the PE in  $q_2$ , since A5 is not available in  $q_2$ , and so cannot rule out compensation. Therefore, the agent must, to act morally, adopt VO4 or VO5, and compensate, even if the state turns out to be  $q_3$ .

In  $q_0/q_1$ , we have two PEs for Hal using VO1: from  $q_0$   $\{A2, A6, A7, A8, A10, A11, A12\}$  and from  $q_1$   $\{A1, A4, A5, A7, A8, A10, A11, A12, A13\}$ . Here A3 is always rejected, reflecting the fact that compensation is impossible. Hal must, however, still choose whether to take the insulin or not. This means that he must adopt a value order which either includes A1 in the PE for both  $q_0$  and  $q_1$ , or which excludes it from both. A1 can be included in both given the preference  $L_H > L_C$ . A1 can, however, only be excluded from the PE for  $q_1$  if  $F_C > L_H$ . VO4 does not decide the issue: thus Hal must choose between self-interest (VO2) and being sacrificial (VO5). Neither choice will be sure to be consistent with morally correct behaviour: VO2 will be wrong in  $q_0$  and V5 will be wrong in  $q_1$ , where the sacrifice is an unnecessary waste. It is because it is unreasonable to require an agent to adopt VO5 (for Carla to expect Hal to do this would require her to adopt the selfish order VO1), that we say that it is morally excusable for Hal to take the insulin in  $q_0/q_1$ .

<sup>4</sup> Note that although we consider only five such value orderings here, there are other orderings that are also possible. For example, an agent may value Hal and Carla's lives equally, but then prefer the freedom of one agent over that of another, e.g.  $(L_H = L_C) > F_H > F_C$ .

The above discussion suggests the following. An agent must consider the PEs relating to every state which it may be in. An action is justified only if it appears in every PE formed using a given value order.

- If VO1 justifies an action, that action is morally obligatory.
- If VO1 does not produce a justified action, then an action justified under VO2, VO4 or VO5 is morally permissible.
- If an action is justified only under VO3, then that action is prudentially correct, but not morally permissible.

Amongst the morally permissible actions we may discriminate according to the degree of preference given to the agent's own interests and we might say: VO2 gives actions which are morally *excusable*, VO4 gives actions which are morally *praiseworthy*, and VO5 gives actions which are *supererogatory* [19], beyond the normal requirements of morality.

These are our suggestions for possible moral nuances, but let us consider again McNamara's concept of supererogation in terms of our model. He envisages a situation in which a person is obliged to see to it that  $p$ , but can meet this obligation in several ways, for example either by doing Action<sub>1</sub> or Action<sub>2</sub>, and that the consequences additional to realising  $p$  will be different if Action<sub>1</sub> rather than Action<sub>2</sub> is performed. In terms of an AATS, from the initial state  $q_0$ , the agent can perform Action<sub>1</sub> to reach  $q_1$  or Action<sub>2</sub> to reach  $q_2$  and in both  $q_1$  and  $q_2$   $p$  is satisfied. The agent will now need to choose between Action<sub>1</sub> and Action<sub>2</sub> on grounds other than  $p$ . It may well be that one state, say  $q_1$ , will enhance the aggregate good of all the agents involved more than  $q_2$ , whereas  $q_2$  will enhance the individual good of the agent making the choice more than will  $q_1$ . The minimum that morality demands in this case will be Action<sub>2</sub>, to reach  $q_2$ , but the supererogatory act will be Action<sub>1</sub> to reach  $q_1$ . In terms of value order, Action<sub>1</sub> will be selected if it conforms with the morally correct ordering, but favours the interests of others over the agent itself, that is VO4, the ordering of the noble agent. Action<sub>2</sub> will be selected if the agent adopts VO2, the self-interested position. On this reasoning, McNamara's supererogatory turns out to be what we have called "morally praiseworthy" here. We wish to draw a distinction between morally praiseworthy, which does correspond to doing more than the minimum that morality requires, and supererogation, since taken too far, consideration of the interests of others may lead to a state in which  $p$  is not satisfied. Here the requirements of morality are not met, albeit that the agent who suffers as a result has chosen to do so.<sup>5</sup>

We may now briefly consider what might be appropriate legislation to govern the situation. We will assume that the following principle governs just laws: that citizens are treated equally under the law. This in turn means that the legislator can only use VO1, as any other ordering requires the ability to discriminate between the interests of the agents involved. We will also assume that the legislator is attempting to ensure that the best outcome (with regard to the interests of all agents) is reached from any given situation. Thus in our example, from  $q_0$  the legislature will be indifferent between  $q_5$  and  $q_{12}$ ; from  $q_1$  and  $q_2$  they will wish to reach  $q_{14}$ ; and from  $q_3$  they will be indifferent between  $q_{16}$  and  $q_{17}$ . Now consider the following possible laws:

**Law 1.** Any agent in Hal's position should be obliged to take the insulin absolutely. This may fail to lead to  $q_{14}$  if such an agent does not compensate in  $q_2$ , and so may not achieve the desired ends. Moreover, in  $q_0$  this requires that  $q_{12}$  rather than  $q_5$  be reached, which prefers the interests of agents in Hal's position to agents in Carla's position.

**Law 2.** Any agent in Hal's position is forbidden to take the insulin unless he pays compensation. This fails to produce the desired outcome in  $q_1$ , where it leads to  $q_7$ .

**Law 3.** Any agent in Hal's position is permitted to take the insulin, but is obliged to compensate if he is able to. This will reach a desired outcome in all states, and is even-handed between agents in Hal and Carla's positions in  $q_0$ . In  $q_3$ , however, it favours the interests of agents in Carla's position over agents in Hal's position by determining which of the two agents ends up with money.

**Law 4.** Any agent in Hal's position is obliged to take the insulin and obliged to compensate if able to. This will reach a desired state in every case, but favours agents in Hal's position in  $q_0$  and agents in Carla's position in  $q_3$ .

<sup>5</sup> There is a saying common amongst Church of England clergy "God preserve us from saints" which also recognises that good intentions can have undesirable consequences, and that self sacrifice is not always the attitude to be enjoined.



Thus if we wish to stick with the principle of not favouring the interests of either agent, we can only *permit* Hal to take the insulin and *permit* Hal to pay compensation: none of the proposed laws are at once even-handed and desirable in all of the possible situations. Under this regime we have no problem in  $q_0$ : the states reached are of equal value, and it is Hal, not the State, who chooses whose interest will be favoured. In  $q_1$  we will reach a desired state provided Hal is not sacrificial. In  $q_2$  we must rely on Hal not being selfish, and acting in a moral fashion. Finally in  $q_3$  we reach a desired state and again Hal chooses whose interests will be favoured. Provided that we can expect agents to act in a morally acceptable, but not supererogatory, fashion, and so use VO2 or VO4, the desired outcomes will be reached. It may be, however, that the legislature will take the view that favouring Carla in  $q_3$  is a price worth paying to prevent selfish behaviour on the part of Hal in  $q_2$ , and pass Law 3. This is a political decision, turning on whether the agents are trusted enough to be given freedom to choose and the moral responsibility that goes with such freedom. A very controlling legislature might even pass Law 4, which gives the agents no freedom of choice, but which reaches a desired state even when agents act purely in consideration of their own interests.

Finally we return to the initial observations of Hare: is it possible to crystallise our reasoning into “not too specific and not too detailed form”? What moral principle might Hal form? First, moral principles which apply to particular states would be too specific. In practice Hal would never have sufficient knowledge of his situation to know which principle to apply. On the other hand, to frame a principle to cover all four initial states would be arguably too general as it would ignore pertinent information. In states  $q_2/q_3$ , the appropriate moral principle is to take and compensate: this ensures that moral censure is avoided, and although it may be, if the state turns out to be  $q_3$ , that Carla’s interests are favoured, Hal is free to make this choice, even if we believe that the State should not impose it. In  $q_0/q_1$ , the choice is not so clear: since moral correctness cannot be ensured, either taking or not taking the insulin is allowed. While taking the insulin is morally excusable, and so an acceptable principle, Hal is free to favour Carla’s interests over his own, provided that it is *his own choice* to do so. While Hal cannot be compelled, or even expected, to be sacrificial, he cannot be morally obliged to be self-interested either.

## 6. Simulations

The approach described above has formed the basis of two simulations, reported in [10] and [11]. The insulin scenario as previously described is realised in [11], while [10] introduces the State as a third agent able to intervene to realise its own values.

The simulation in [11] required, as noted at the beginning of Section 2, some refinement to the representation in order to block some undesirable arguments. First it was necessary to distinguish situations in which insulin was available for purchase, to prevent Hal from simply buying insulin to replace what he had lost. We modelled this as a single Boolean, indicating whether the chemist was open or not. Second it was necessary to distinguish situations where the agent was in critical need of insulin from those where it had more time to find insulin. This was to prevent Carla from dying when compensated by Hal before she was able to buy insulin. To this end the variable for life was able to take the values 2 (good health), 1 (critical) and 0 (dead). Finally we needed to give some impact to compensating Carla when she already had money. This was achieved by making the variable for money also take the values 2 (more than enough), 1 (enough to buy insulin) and 0 (not enough to buy insulin). Although not changing the principles of the approach, this made the automata more complicated and the state space much larger: the worth of computer simulation on this increased size of problem was evident.

Once these adjustments had been made, the simulation confirmed the reasoning given above. One interesting point was that where two agents both had money, neither agent did anything until Carla’s health became critical: Carla waited to be compensated, and Hal was not prepared to compensate. In the next move Carla would buy insulin and Hal would compensate her only if noble or sacrificial (in the simulations the agents were aware of each other’s financial state).

The simulation, of course, also confirmed that where neither had money one of them died. The second simulation [10] introduced the State modelled as a third agent. The discussion of possible laws to influence the behaviour of the agents in Section 5 conceived of the State as a passive ring holder, issuing laws, but not intervening to change the situation. This conception is rather like the “Social Laws” approach to co-ordination in multi-agents systems which originated in work such as [18] and which forms the basis of [22] and [20]. In this approach the idea is to form social laws which, if complied with by the agents, will ensure that certain undesired states are not reached. It is argued that agents will comply with these laws since they will see that their interests are served by avoiding these states. This

Table 3  
Actions for the Different Situations

Preference	$M_C = 0$		$M_C > 0$	
	$M_H = 0$	$M_H > 0$	$M_H = 0$	$M_H > 0$
$L > F_S > F$	give insulin	fine + give insulin	give insulin	fine + give insulin
$L > F > F_S$	give insulin	give insulin	give insulin	give insulin
$F_S > L > F$	nothing	fine	buy insulin	fine + buy insulin

compliance is considered likely even if the law is asymmetric in that it benefits one agent more than another, since the less favoured agent nonetheless receives the benefits of the assurance that the undesired states are avoided. No such mutually advantageous social law is possible in our example, since from some initial states an undesired state cannot be avoided. Moreover, although an undesirable state must inevitably be reached, the choices of the agents determine which of them will suffer: our argument was that this decision should be made by the agents rather than the State.

In practice, of course, the State is not entirely passive, but can intervene so as to affect the states reached, either by imposing sanctions, or by providing benefits. In the second simulation [10], we modelled the effect of allowing the State to intervene so that it was able to ameliorate the undesired consequences. The State was given two actions: it could supply an agent with insulin (which it would do if the agent had no insulin) and it could fine an agent who took another's insulin and did not compensate (provided the fine could be paid). With these actions available it is possible, should the State see it as in its interests to do so, to ensure that all the agents survive, whatever the initial state. In order to determine the rational response of the State in particular situations, that State is ascribed values of its own so that it can use practical arguments in the same way as the other agents. The State's values were the life of its citizens, its own resources, and the resources of its citizens.

A sample of the results given in Table 3 shows the results for two selfish agents each of which may or may not have money, and three value orderings for the State.

Broadly these results were as we would expect: the State will give Carla insulin, provided it prefers the lives of its citizens to its resources, and will fine Hal if it can, unless it values its citizen's resources over its own. There are, however, some interesting observations.

One is that Carla will never buy insulin, even if she can, unless she believes the State to value its resources over the lives of its citizens. If she believes the State will give her insulin, she will take the opportunity to save her money. This situation was paralleled by prescription charges in the UK. Originally prescriptions for required medicines were free under the UK National Health Service, but over time it was seen as anomalous that those who could pay for their medicines did not do so. As a result prescription charges were introduced, broadly for those in work. Note that the simulation would have drawn the attention of a policy maker to this feature.

A second point is that Hal *never* compensates Carla if the State will give her insulin, unless he is noble or sacrificial. Since Hal is aware that Carla's life is not threatened, he does not voluntarily give up his money. This is not, however, what we would wish to happen. Two possibilities offer themselves: the State could make fines punitive, so that it is in Hal's financial interests to compensate voluntarily. This, however, works only where Hal has more than enough money. An alternative would be to introduce an extra value to represent the social stigma of being fined. Where Hal gives sufficient regard to this value, even a selfish Hal will compensate Carla. This indicates a distinct benefit for a society in which agents have this kind of respect for (or fear of) the law.

A third point is the importance of CQ17, which requires agents to consider what they expect other agents to do, in this second simulation. This in turn requires the agents to have knowledge of, or make assumptions about, the value priorities of the other agents. In the case of the State, this should not be a problem, since it will promulgate its legislation and so declare its intentions in advance. There could, however, be a problem if the actions of the State were subject to some form of discretion, as this would mean that the agent could not base its decisions on a reliable view of how the State will act. If the giving of insulin were discretionary, for example, Carla would be somewhat foolhardy not to buy it herself if she could. This uncertainty could well frustrate the intentions of the legislation. With regard to other agents, the safest course of action is simply to assume that the other agents are selfish. Thus the State will give Carla insulin, even when Hal could compensate her, since it cannot rely on Hal doing "the right thing".

This assumption is probably the best that can be made when dealing with anonymous or unknown agents, or when the State is considering how agents will behave in relation to itself. Thus the State could probably rely on Carla to buy insulin if she were able and the State did not supply it, but would be unwise to rely on Hal compensating Carla in

the case where he had money and she did not. In practice, however, we often do need to rely on other agents acting in a less than entirely selfish way, without expecting them to go beyond what morality demands. Working, family and social relationships would become impossible without sometimes acting in the belief that another agent will give some consideration to our interests. In these particular situations knowledge of the individuals can give rise to this trust: the advantage of a common moral code taken for granted throughout a society is that we can make similar reliance on people of whom we have no particular knowledge.

A further point of interest is that where there was no State intervention, sometimes bad things (someone dying) were inevitable, but if not inevitable could be avoided by a degree of unselfishness on the part of the agents. State intervention, at least in the form of the actions modelled here, ensures that bad things do not happen, but no longer requires any unselfishness on the part of the agents: Hal is not motivated to compensate Carla since she is removed from danger, and can only be *coerced* to compensate by threat of sanctions. Viewed in this light, the well-intentioned State militates against the development of a moral sense. If, however, we motivate Hal to avoid being fined by introducing an additional value representing respect for law, we can produce the unselfish behaviour once more. Here respect for law is substituting for the notion of morally reasonable behaviour that arose in the ungoverned situation.

## 7. Concluding remarks

In this paper we have described how agents can reason about what they ought to do in particular situations and how moral principles can emerge from this reasoning. An important feature is how their choices are affected by the degree of consideration given to the interests of the other agents involved in the situation, which is captured by an ordering on the values used to ground the relevant arguments. Different value orders will attract varying degrees of moral praise and censure. We have further illustrated this by reference to some computer simulations, including a model where the State is able to intervene. Where the State intervenes the proper behaviour required of the agents is altered.

In future work we will wish to consider further the relation between the various kinds of “ought” we have identified here. In particular, it might be conjectured that reasoning with the morally reasonable value orders VO2 and VO4 will always lead to an outcome which is desirable when aggregating the interests of the agents involved. The simulations empirically support this conjecture, but analytic investigation would be more rigorous. A second interesting line of inquiry would be to increase the number of agents involved and to use more complicated scenarios, to consider the effect of agents having different attitudes towards the others depending on their inter-relationships, modelling notions such as kinship, community and national groupings. Such investigations on complicated state spaces will undoubtedly benefit from support from computer simulation and model checking techniques such as those used to explore social laws, as in [20]. We would hope that this line of inquiry would yield insight into the emergence of norms of cooperation.

## Acknowledgements

This paper is a revised and extended version of a paper presented at the Eighth International Workshop on Deontic Logic in Computer Science (DEON 2006). We thank the anonymous reviewers and the participants of that workshop for their comments and suggestions. We would also like to thank Alison Chorley for her contributions to our work on modelling practical reasoning. We also acknowledge partial support received from the European Commission through Project ASPIC (IST-FP6-002307).

## References

- [1] R. Alur, T.A. Henzinger, O. Kupferman, Alternating-time temporal logic, *ACM Journal* 49 (5) (2002) 672–713.
- [2] L. Amgoud, S. Kaci, On the generation of bipolar goals in argumentation-based negotiation, in: I. Rahwan, P. Moraitis, C. Reed (Eds.), *Proceedings of the First International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2004)*, in: *Lecture Notes in Artificial Intelligence*, vol. 3366, Springer, 2004, pp. 192–207.
- [3] K. Atkinson, What should we do?: Computational representation of persuasive argument in practical reasoning, PhD thesis, Department of Computer Science, University of Liverpool, Liverpool, UK, 2005.
- [4] K. Atkinson, T. Bench-Capon, Addressing moral problems through practical reasoning, in: L. Goble, J.Ch. Meyer (Eds.), *Deontic Logic and Artificial Normative Systems*, in: *Lecture Notes in Artificial Intelligence*, vol. 4048, Springer, 2006, pp. 8–23.
- [5] K. Atkinson, T. Bench-Capon, Action-based alternating transition systems for arguments about action, in: *Proceedings of the Twenty Second Conference on Artificial Intelligence (AAAI 2007)*, AAAI Press, 2007, pp. 24–29.



- [6] K. Atkinson, T. Bench-Capon, Argumentation and standards of proof, in: *Proceedings of the Eleventh International Conference on AI and Law (ICAAIL 2007)*, ACM Press, 2007, pp. 107–116.
- [7] K. Atkinson, T. Bench-Capon, P. McBurney, A dialogue game protocol for multi-agent argument for proposals over action, *Autonomous Agents and Multi-Agent Systems* 11 (2) (2005) 153–171.
- [8] K. Atkinson, T. Bench-Capon, P. McBurney, Computational representation of practical argument, *Synthese* 152 (2) (2006) 157–206.
- [9] T. Bench-Capon, Persuasion in practical argument using value based argumentation frameworks, *Journal of Logic and Computation* 13 (3) (2003) 429–448.
- [10] A. Chorley, T. Bench-Capon, Modelling state intervention in cases of conflicting interests, in: T. van Engers (Ed.), *Legal Knowledge and Information Systems, JURIX 2006: The Annual Conference*, in: *Frontiers in Artificial Intelligence and Applications*, vol. 152, IOS Press, 2006, pp. 31–40.
- [11] A. Chorley, T. Bench-Capon, P. McBurney, Automating argumentation for deliberation in cases of conflict of interest, in: P.E. Dunne, T.J.M. Bench-Capon (Eds.), *Computational Models of Natural Argument, Proceedings of COMMA 2006*, in: *Frontiers in Artificial Intelligence and Applications*, vol. 144, IOS Press, 2006, pp. 279–290.
- [12] C.G. Christie, *The Notion of an Ideal Audience in Legal Argument*, Kluwer Academic Press, 2000.
- [13] J. Coleman, *Risks and Wrongs*, Cambridge University Press, 1992.
- [14] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77 (1995) 321–357.
- [15] S.O. Hansson, *The Structure of Values and Norms*, Cambridge University Press, 2001.
- [16] R.M. Hare, *Freedom and Reason*, The Clarendon Press, Oxford, UK, 1963.
- [17] P. McNamara, Doing well enough: Towards a logic for common sense morality, *Studia Logica* 57 (1996) 167–192.
- [18] Y. Moses, M. Tennenholtz, Artificial social systems, *Computers and AI* 14 (6) (1995) 533–562.
- [19] J.O. Urmson, Saints and heroes, in: A.I. Melden (Ed.), *Essays in Moral Philosophy*, University of Washington Press, 1958, pp. 198–216.
- [20] W. van der Hoek, M. Roberts, M. Wooldridge, Social laws in alternating time: Effectiveness, feasibility and synthesis, *Synthese* 156 (1) (2007) 1–19.
- [21] D.N. Walton, *Argumentation Schemes for Presumptive Reasoning*, Lawrence Erlbaum Associates, Mahwah, NJ, 1996.
- [22] M. Wooldridge, W. van der Hoek, On obligations and normative ability: Towards a logical analysis of the social contract, *Journal of Applied Logic* 3 (2005) 396–420.